DOCUMENT RESUME

ED 303 481 TM 012 565

AUTHOR Perkins, Kyle; Duncan, Ann
TITLE An Internal Construct Validation Study of the "Iowa Tests of Basic Skills" (Level 12, Form G) Reading Comprehension Test Items.
PUB DATE 87
NOTE 24p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Construct Validity; Elementary School Students; Functional Reading; Generalization; *Grade 6; Inferences; Intermediate Grades; *Reading Comprehension; Reading Skills; Reading Tests; Test Items; *Test Validity
IDENTIFIERS *Internal Validity; *Iowa Tests of Basic Skills

ABSTRACT

An assessment analysis was performed to determine whether sets of items designed to measure three different subskills of reading comprehension of the Iowa Tests of Basic Skills (ITBSs) did, in fact, distinguish among these subskills. The three major skills objectives were: (1) facts; (2) generalizations; and (3) inferences. Data from administration of the ITBSs to 82 sixth-graders in the Carterville (Illinois) Grade School were analyzed. The reading comprehension test was divided for the purpose of analysis into: (1) 17 items assessing facts; (2) 22 items assessing generalizations; and (3) 17 items assessing inferences. Internal construct validity was calculated with a procedure developed by G. H. Henning (1987). Results suggest that, for this sample, the reading comprehension items did not have robust coefficients of internal construct validity. Only 17 of the 56 items exhibited internal construct validity when compared with two criterion measures. Students' strengths and weaknesses in the different comprehension skills could not be determined, because there was a lack of items demonstrating robust discrimination among these three reading comprehension skills. It is concluded that school systems mandated to show that their reading instructional programs effect mastery in different reading comprehension skills will have a difficult task of offering empirical proof if the ITBSs are used to elicit the test data. Six tables contain study data and an appendix lists skills objectives. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED303481

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

KYLE PERKINS

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

An Internal Construct Validation Study of the Iowa Tests of Basic Skills (Level 12, Form G) Reading Comprehension Test Items

Kyle Perkins
Southern Illinois University

and

Ann Duncan
Carterville Grade School

TM 012 565

Abstract

This paper reports the results of an assessment analysis of the Iowa Tests of Basic Skills (Level 12, Form G) reading comprehension test items to determine whether the sets of items designed to measure three different subskills distinguished between the subskills which they were intended to measure. The results indicated that, in general, the three sets of items appeared to be alternative measures of the same underlying construct.

An Internal Construct Validation Study of the Iowa Tests of Basic Skills (Level 12, Form G) Reading Comprehension Test Items

INTRODUCTION

Since the mid-1940s there has been little agreement in the reading-research literature about the number and the identity of factors involved in reading comprehension. The principal theme in this research area has been a search for the correlates of overall reading skill. Davis (1944) administered a battery of reading comprehension subskill tests and, using factor-analytic techniques, attempted to determine the number of distinctive factors involved in reading comprehension. He reported finding five factors: knowledge of word meaning; reasoning in reading; concentration on literal sense meaning; following the structure of a passage; and recognizing the writer's mood and literary techniques. Of these five factors, Davis claimed that memory for word meaning and reasoning in reading were the two most important factors in reading comprehension. In essence, Davis concluded that reading comprehension is not a unitary process.

Thurstone (1946), using a different factor-analytic approach, reanalyzed Davis's data. The results of that reanalysis have been interpreted in several ways. One interpretation is that these data reveal only one robust factor, while another interpretation yields three major factors: ability to draw inferences; ability to grasp the main idea of a paragraph; and vocabulary knowledge. Following the studies of Davis and Thurstone, a tradition of factor analytic studies developed which identified a reading comprehension factor (Botzum, 1951: Clark, 1972; Fruchter, 1948; Wrigley, Saunders, and Newhaus, 1958).

During the interim between the Davis-Thurstone studies and the Botzum

study, a most ambitious study to determine the number of distinctive factors involved in reading comprehension was conducted by Holmes and his successor, Singer. In what has come to be known as the substrata factor theory of reading, Holmes and Singer (1961) selected a set of eighty candidate variables to study and from that set of variables they attempted to identify the most significant causal factors which account for attained reading comprehension proficiency. Two principal components were identified: power and speed. Holmes and Singer noted that a multiplicity of skills and processes underlie these two components and suggested that the reader organizes these skills and processes into a working system according to the demands of the reading task. The substrata factor in this analysis spoke of printed stimuli being transformed into a word-recognition substratum and the stimuli being transformed into meaning. These two types of transformations are said to be synchromesed into a harmonious working system, and reasoning in context, word meaning, and word recognition are entailed.

Researchers have reached no consensus about whether reading comprehension is a unitary or a divisible competence. There are several reasons for this general lack of agreement. First, there is disagreement in the literature about the number of factors underlying reading comprehension. For example, results have been published claiming that reading is a holistic process (Thorndike, 1974). Others have offered experimental findings which support a skills or components approach (Davis, 1972; Spearritt, 1972). There is variation in both the type and the number of factors identified. To add to the confusion, proponents of the single factor (holistic) hypothesis have utilized the same analytical techniques employed by the divisible competence theorists to demonstrate single factor solutions (Clark, 1972; Thorndike, 1974). In fact, Thorndike claimed that reading is reasoning and that it cannot be further decomposed into separate components.

Further, other researchers have used logical, non-factor analyses to derive reading skills with a similar lack of concordance (Baker and Stein, 1978; Rystrom, 1970; Vernon, 1962).

Three recent studies merit special consideration here: Drahozal and Hanna (1978); Haertel (1980), and Farr and Carey (1986). Drahozal and Hanna identified three tasks (literal, translation, and higher-order) assessed in the Nelson Reading Skills Tests (Hanna, Schell, and Schreiner, 1977) for grades three through nine and computed intercorrelation matrices corrected for attenuation between the tasks in order to estimate the extent of the relationship between the students' "true" test scores. Drahozal and Hanna reported that their results indicated that the three different tasks appeared to be different means of measuring the same underlying construct, that is, no evidence could be found to show that the three different tasks reflected three different independent constructs.

When Haertel examined the test items and the item responses from standardized reading comprehension tests published prior to 1970 in terms of nine comprehension skills, he found no evidence suggesting uniqueness associated with the subskills. Rather, he found that reading comprehension appeared to be a "single dichotomous skill"--a reader can or cannot comprehend. Haertel's research and other criticisms of reading comprehension assessment instruments lead Farr and Carey to suggest a series of assessment analyses, one of which was, "Do reading assessment instruments now in use distinguish between the subskills that they profess to measure?" (p. 52).

The present study

The purpose of this paper is to report the results of an internal construct validation study of the items designed to measure three major

skills objectives in the Iowa Tests of Basic Skills (Level 12, Form G) reading comprehension test (Hieronymus, Hoover, and Lindquist, 1986). The three major skills objectives are (1) facts, (2) generalizations, and (3) inferences. The task at hand was to determine if the fact questions measured a construct which was different from the constructs measured by the generalization questions and the inference questions. In other words, could evidence be found to show that the sets of questions designed to measure the three different major skills objectives did indeed measure something unique? Another way of stating the research question is to ask whether the three sets of questions designed to measure the three major skills objectives demonstrated discriminant validity. This piece of research was intended to answer Farr and Carey's question with regard to the different sets of reading comprehension items in the Iowa Tests of Basic Skills (Level 12, Form G).

METHOD

Subjects

The subjects who generated the data for this research were eighty two sixth-graders enrolled in the Carterville (Illinois) Grade School. The Iowa Tests of Basic Skills complete battery is administered each April as a part of the school's assessment program.

Instrumentation

The Preliminary Teacher's Guide (Hieronymus, Hoover, and Lindquist, 1986) described the skills objectives as follows:

> The sixteen skills objectives represented in the Reading Comprehension test are classified into three major divisions: facts-to recognize and understand stated factual details and relationships; generalizations-to develop generalizations from a selection; inferences-to infer underlying relationships. These three categories correspond closely to those of "literal meaning," "interpretative meaning," and "evaluative meaning" used by some authorities in classifying objectives of reading.
>
> Objectives involving factual details and relationships are often represented by items of the who-what-when-where type. The first three skills objectives--description, categorization, and relationships--correspond to the descriptive, categorical, and relational cognitive style modes of information processing described by Kagan, Moss and Sigel (1963) and by Robinson and Gray (1974). The fourth objective is concerned with deducing the special meaning of words and phrases from context.
>
> The second major class of skills objectives, inferences, is represented for the most part by why? questions. The four objectives are understanding cause-and-effect relationships, drawing conclusions, inferring traits and feelings, and inferring motives and reasons for the actions of characters.
>
> Objectives in the third category, generalizations, involve higher order skills concerned with main ideas, organization, application, purpose and viewpoint of the author, and the structure, mood, or style of the selection (pp. 44-45).

(Put Appendix A about here)

Appendix A presents the skills objectives assessed in the Level 12, Form G Iowa Tests of Basic Skills reading comprehension test and the number of items constructed to measure each skill objective.

Procedure

For purposes of analysis, the reading comprehension test was partitioned into three subtests: (1) 17 items assessing facts; (2) 22 items assessing generalizations; and (3) 17 items assessing inferences. For each person, the following person scores were calculated: (1) a fact total score; (2) a generalization total score; and (3) an inference total score.

(Put Table 1 about here)

Table 1 presents the means, standard deviations, and the estimates of reliability for the three subtests and the total test. The Kuder-Richardson Formula 20 method was used to estimate the coefficients of reliability.

The internal construct validation procedure employed in this research was presented in Henning (1987). Henning's procedure makes the following assumptions, in the context of this research. If, for example, the fact items have construct validity, then the point-biserial correlation between each fact item and the fact subtest total scores should be higher than the point-biserial correlations of the fact items with the generalization subtest or with the inference subtest total scores. Following Henning, we can illustrate this concept mathematically as follows:

$$r_{F_1FT} > r_{F_1GT}$$

$$r_{F_2FT} > r_{F_2GT}$$

$$r_{F_3FT} > r_{F_3GT}$$

where F_1 = fact item number 1

FT = fact subtest total

GT = generalization subtest total

These symbols indicate that correlations of individual fact items with their own subtest totals should be greater than the correlations of the fact items with the generalization subtest totals.

The purpose of internal construct validation is to provide evidence that the underlying theoretical constructs being measured are valid <u>per se</u>. The procedure involves calculating correlations among items or subscales

of a test; consequently, the validation criteria are drawn from the test battery itself, and it is not necessary to administer concurrent tests to serve as validation criteria.

For each item a point-biserial correlation was calculated between the success or failure on each item and the total scores on each of the three subtests. Point-biserial correlations involve one dichotomous variable (pass-or-fail) and a continuous variable (the subtest total score). The point-biserial correlation is so named because there are two series of persons being observed on the Y variable: (1) those who failed an item and scored 0 on the dichotomous measure, X and (2) those who passed the item and scored 1 on X, the dichotomous measure.

When an item is correlated with its own total score, the resulting point-biserial correlation coefficient is spuriously inflated because the presence of the item in the total score is responsible for part of the coefficient's magnitude. Therefore, all item correlations with their own total score were corrected for part-whole overlap to remove the contribution of the item to the total score.

(Put Tables 2, 3, and 4 about here)

Tables 2, 3, and 4 present the point-biserial correlations for each of the three sets of items. The alpha-numerical codes in parentheses refer to the identification of the different skills objectives presented in Appendix A.

To answer the research question--do the items distinguish between the subskills that they profess to measure--two different analyses were conducted. First, for each set of questions designed to measure a major skill objective, we determined the proportion of valid items within that

subtest. In this context, valid means that an item correlated more highly with its own subtest total than with the total from another subtest. In this manner, we were able to report the construct validity of each subtest as a single coefficient. Following Henning, we refer to this coefficient as the internal construct validity proportion.

For the second analysis, we used the Fisher Z transformation to transform the point-biserial correlation coefficients in Tables 2, 3, and 4 from a distorted ordinal scale to a normal interval scale. The resulting Z scores were submitted to one-factor analysis of variance. Separate one-factor analyses of variance were conducted for the data in Tables 2, 3, and 4. If each set of items designed to measure a specific major skill objective has internal construct validity, then the mean correlation, for example, for the fact-fact total column should be greater than the mean correlations for the fact-generalization total column and for the fact-inference total column. As an example for Table 2, the two hypotheses to be tested can be stated as follows:

Null hypothesis: $\bar{X}$ F-FT = $\bar{X}$ F-GT = $\bar{X}$ F-IT

Alternative hypothesis: $\bar{X}$ F-FT $\neq$ $\bar{X}$ F-GT $\neq$ $\bar{X}$ F-IT

(The mean point-biserial correlations shown at the bottom of each column in Tables 2, 3, and 4 are the result of averaging the Fisher Z scores and transforming each average Z score back to a point-biserial correlation coefficient).

RESULTS AND DISCUSSION

(Put Table 5 about here)

Table 5 presents the results of the first analysis. For purposes of comparison we have included for each subtest its maximum possible validity in the most ideal circumstances which is the square root of the reliability of the subtest, or $\sqrt{r_{tt(x)}}$ (Henning, 1987, p. 90). This method of estimating the maximum possible validity of a test reflects the known relationship between reliability and validity, for a test cannot be valid unless it is first reliable, but a test can be reliable, but not be valid for a specific educational use.

The validity coefficients want explication. These coefficients represent the proportion of internally valid items when compared with the other two subtest totals. We offer two examples. First, an inspection of Table 2 indicates that seven fact items had fact item-fact subtest total point-biserial correlations corrected for part-whole overlap higher than the fact item-generalization subtest total point-biserial correlations. Those seven items were items 46, 49, 71, 72, 74, 90, and 91; thus, 7/17 = .412. Second, an inspection of Table 3 indicates that thirteen generalization items had higher generalization item-generalization subtest total point-biserial correlations corrected for part-whole overlap than generalization item-fact subtest total point-biserial correlations. Those item numbers were 54, 58, 68, 75, 76, 77, 79, 81, 83, 84, 92, 93, and 94; thus, 13/22 = .591.

The results of the first analysis suggest two observations to us. First, the internal construct validity coefficients for each subtest fall far short of their maximum possible validity coefficients. Second, these results suggest an opportunity for economy of measurement along the following

lines. There were only five fact items whose fact-fact total point-biserial correlations were larger than both the fact-generalization total and fact-inference total point-biserial correlations: 46, 49, 71, 74, and 91, and this count ignores the fact that the correlation for item number 46 was not significant at the .05 level. In other words, for this sample of students only five of the seventeen fact items demonstrated discriminant or internal construct validity (5/17 = .294).

Seven generalization items had generalization-generalization total point-biserial correlations corrected for part-whole overlap greater than both the generalization-fact total correlations and the generalization-inference total correlations: 58, 68, 75, 77, 79, 81, and 84 (7/22 = .318). And five inference items demonstrated discriminant validity: 52, 56, 87, 97, and 98 (5/17 = .294).

In summary, only 17 of the 56 items (or 30.36 percent of the reading comprehension items) exhibited internal construct validity when they were compared with both criterion measures. If our purpose is to assess how well our students recognize facts, then only five items generated unique variance. It does not seem warranted to use a fact item to give us information about how well our students recognize and comprehend factual information when that particular item correlates higher with the inference total than it correlates with its own fact subtest total.

(Put Table 6 about here)

Table 6 presents the summary tables for the three one-factor analyses of variance. In all three analyses, the null hypothesis was retained; there were no significant differences between the columns of Fisher Z scores for any of the three sets of items; therefore, there was no reason to conduct

a post hoc multiple comparison analysis such as the Scheffé technique or the Newman-Keuls method on the means.

CONCLUSION

The results of this study suggest that for this particular sample of students the reading comprehension items in the Iowa Tests of Basic Skills (Level 12, Form G), in general, do not exhibit robust coefficients of internal construct validity. An examination of both the test and the teacher's manual suggests that the test contains questions which appear to assess different skills. The test user and the teacher could benefit from subscore test reports in order to enchance instructional practices if such subscores could identify the students' relative strengths and weaknesses in the various reading comprehension skills. The results from this study indicate that it is not possible to identify students' strengths and weaknesses in the different comprehension skills because there was a paucity of items which demonstrated robust discrimination between three reading comprehension skills purportedly assessed by the Iowa battery. And finally, we note that if a school system were mandated to demonstrate that its reading instructional program effects mastery in different reading comprehension skills, that school system will have an exceedingly difficult task of offering empirical proof, if the Iowa Tests of Basic Skills is the assessment instrument which is utilized to elicit the test data.

REFERENCES

Baker, L., & Stein, N. L. (1978). The development of prose comprehension skills (Technical Report No. 188). Urbana, IL: University of Illinois, Center for the Study of Reading.

Botzum, W. A. (1951). A factorial study of reasoning and closure factors. Psychometrika, 16, 361-386.

Clark, N. L. (1972). Hierarchial structure of comprehension skills. Hawthorn, Victoria, Australia: ACER.

Davis, F. B. (1944). Fundamental factors of comprehension in reading. Psychometrika, 9, 185-197.

Davis, F. B. (1972). Psychometric research on comprehension in reading. Reading Research Quarterly, 7, 628-678.

Drahozal, E. C., & Hanna, G. S. (1978). Reading comprehension subscores: Pretty bottles for ordinary wine. Journal of Reading, 21, 416-420.

Farr, R., & Carey, R. F. (1986). Reading: What can be measured?, second edition. Newark, DE: International Reading Association.

Fruchter, B. (1948). The nature of verbal fluency. Educational and Psychological Measurement, 8, 33-47.

Haertel, E. (1980). A study of domain heterogeneity and content acquisition. Evanston, IL: Cemrel.

Hanna, G. S., Schell, L. M., & Schreiner, R. (1977). The Nelson reading skills test. Boston: Houghton Mifflin.

Henning, G. H. (1987). A guide to language testing: Development, evaluation, research. Cambridge, MA: Newbury House Publishers.

Hieronymus, A. N., Hoover, H.D., & Lindquist, E. F. (1986). Iowa tests of basic skills (Form G, Level 12). Chicago: The Riverside Publishing Company.

Hieronymus, A. N., Hoover, H. D., & Lindquist, E. F. (1986). Preliminary teacher's guide. Multilevel battery levels 9-14. Chicago: The Riverside Publishing Company.

Holmes, J. A., & Singer, H. (1961). The substrata-factor theory: The substrata factor differences underlying reading ability in known groups. Final report covering contracts 538 and 538A. Washington, D. C.: Office of Education, U. S. Department of Health, Education, and Welfare.

Kagan, J., Moss, H. A., & Sigel, I. E. (1963). Psychological significance of styles of conceptualization. Society for Research in Child Development Monographs, 28, 73-112.

Robinson, J. E., & Gray, J. L. (1974). Cognitive style as a variable in school learning. Journal of Educational Psychology, 66, 793-799.

Rystrom, R. (1970). Toward defining comprehension: a second report. Journal of Reading Behavior, 2, 144-157.

Spearritt, D. (1972). Identification of subskills of reading comprehension by maximum likelihood factor analysis. Reading Research Quarterly, 8, 92-111.

Thorndike, R. L. (1974). Reading as reasoning. Reading Research Quarterly, 9, 137-147.

Thurstone, L. (1946). Note on reanalysis of Davis's reading test. Psychometrika, 11, 185-188.

Vernon, P. E. (1962). The determinants of reading comprehension. Ecucational and Psychological Measurement, 22, 269-286.

Wrigley, C., Saunders, D. R., & Newhaus, J. O. (1958). Application of the quatrimax method of rotation to Thurstone's P. M. A. study. Psychometrika, 23, 115-170.

Appendix A
Skills Objectives

SKILLS OBJECTIVES	NUMBER OF ITEMS
F Facts: To Recognize and Understand Stated Factual Details and Relationships (Literal Meaning)	17
F1 Description: To understand factual details relating to description of people, places, objects, and events	5
F2 Categorization: To understand factual details relating to classification	1
F3 Relationships: To understand functional relationships, time, and sequence	6
F4 Contextual Meaning: To deduce the meanings of words or phrases from context	5
I Inferences: To Infer Underlying Relationships (Interpretative Meaning)	17
I1 Cause and Effect: To understand cause, effect, concomitance, and interaction	6
I2 Draw Conclusions: To draw conclusions from information and relationships	5
I3 Traits and Feelings: To infer traits, feelings, and emotions of characters	3
I4 Motives: To infer the motives and reasons for the actions of characters	3
G Generalizations: To Develop Generalizations from a Selection (Evaluative Meaning)	22
G1 Main Idea: To recognize the main idea or topic of a paragraph or selection	6
G2 Organization: To understand the organization of a paragraph or selection	2
G3 Application: To apply information through generalization or prediction	1
G4 Purpose: To recognize the author's purpose, motive, or intention	4
G5 Viewpoint: To recognize the author's viewpoint, attitude, or bias	1
G6 Figurative Language: To interpret figurative language	2
G7 Mood: To recognize the mood or tone of a selection	2
G8 Style: To recognize qualities of style or structure	4

(Hieronymus, Hoover, and Lindquist, 1986, p. 45)

Table 1
Means, standard deviations, and estimates of reliability

Test	No. of examinees	No. of items	Mean no.of correct responses	S.D.	r_{tt}
Fact	82	17	12.39	2.90	.691
Generali-zation	82	22	16.23	3.63	.753
Inference	82	17	11.67	3.05	.708
Total	82	56	40.29	8.73	.888

Table 2
Point-biserial correlations
Fact items

Item no.	(Skill objective)	Fact total	Generali-zation total	Inference total
46	(F1)	.178	.169	.141
47	(F3)	.099	.167	.182
48	(F3)	.330	.370	.228
49	(F4)	.258	.125	.182
55	(F4)	.278	.328	.377
61	(F1)	.093	.136	.153
62	(F4)	.243	.341	.362
70	(F4)	.199	.328	.320
71	(F3)	.337	.301	.172
72	(F3)	.317	.274	.332
74	(F1)	.433	.351	.420
78	(F4)	.027	.126	.033
89	(F1)	.426	.464	.497
90	(F3)	.488	.455	.510
91	(F2)	.492	.303	.450
96	(F1)	.187	.371	.342
99	(F3)	.306	.329	.377
	Mean	.281	.294	.305

Point-biserial correlations of .183 and higher are significant at the .05 level, one-tailed test; point-biserial correlations of .256 and higher are significant at the .01 level, one-tailed test.

Table 3
Point-biserial correlations
Generalization items

Item no.	(Skill objective)	Fact total	Generalization total	Inference total
45	(G1)	.203	.193	.177
50	(G3)	.228	.032	.139
51	(G5)	.431	.350	.357
54	(G4)	-.088	.002	.013
57	(G6)	.261	.249	.188
58	(G1)	.258	.271	.248
65	(G8)	.456	.445	.477
67	(G1)	.392	.264	.274
68	(G4)	.424	.440	.308
75	(G2)	.247	.281	.134
76	(G7)	.217	.227	.280
77	(G4)	.217	.266	.073
79	(G6)	.465	.547	.515
80	(G8)	.411	.386	.412
81	(G8)	.203	.323	.251
83	(G7)	.222	.308	.322
84	(G4)	.222	.348	.328
92	(G1)	.336	.365	.396
93	(G8)	.358	.443	.448
94	(G1)	.339	.396	.472
95	(G2)	.243	.113	.288
100	(G1)	.332	.292	.301
	Mean	.294	.302	.304

Point-biserial correlations of .183 and higher are significant at the .05 level, one-tailed test; point-biserial correlations of .256 and higher are significant at the .01 level, one-tailed test.

Table 4
Point-biserial correlations
Inferenc e items

Item no.	(Skill objective)	Fact total	Generali-zation total	Inference total
52	(I3)	.074	.051	.115
53	(I4)	.014	.118	.089
56	(I3)	.327	.302	.339
59	(I4)	.268	.378	.306
60	(I1)	.348	.226	.284
63	(I2)	.524	.504	.365
64	(I3)	.426	.418	.336
66	(I1)	.225	.166	.028
69	(I1)	.366	.335	.359
73	(I2)	.218	.119	.125
82	(I2)	.260	.270	.299
85	(I4)	.435	.463	.442
86	(I2)	.424	.455	.355
87	(I2)	.318	.318	.332
88	(I1)	.398	.394	.243
97	(I1)	.363	.418	.463
98	(I1)	.173	.231	.306
	Mean	.308	.309	.306

Point-biserial correlations of .183 and higher are significant at the .05 level, one-tailed test; point-biserial correlations of .256 and higher are significant at the .01 leve, one-tailed test.

Table 5
Internal construct validity coefficients

Subtest	Maximum validity possible	Criterion measure		
		1	2	3
1. Fact	.831		.412	.353
2. Generalization	.868	.591		.409
3. Inference	.841	.529	.529	

Table 6
Analysis of variance summary tables

Fact items (Table 2)

Source of variation	SS	d.f.	M.S.	F
Between groups	.0056	2	.0028	.0139
Within groups	.9679	48	.0201	
Total	.9735	50		

Generalization items (Table 3)

Source of variation	SS	df	MS	F
Between groups	.0015	2	.0007	.0381
Within groups	1.2946	63	.0205	
Total	1.2961	65		

Inference items (Table 4)

Source of variation	SS	df	MS	F
Between groups	.0076	2	.0038	.1864
Within groups	.9822	48	.0204	
Total	.9898	50		